Published online 24 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D25–D29***
doi:10.1093/nar/gks1152

DDBJ new system and service refactoring

Osamu Ogasawara[1,*], Jun Mashima[1], Yuichi Kodama[1], Eli Kaminuma[1], Yasukazu Nakamura[1,*], Kousaku Okubo[1] and Toshihisa Takagi[1,2,*]

[1]DDBJ Center, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan and [2]National Bioscience Database Center, Japan Science and Technology Agency, Tokyo 102-8666, Japan

Received September 15, 2012; Revised October 24, 2012; Accepted October 25, 2012

ABSTRACT

The DNA data bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) maintains a primary nucleotide sequence database and provides analytical resources for biological information to researchers. This database content is exchanged with the US National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) within the framework of the International Nucleotide Sequence Database Collaboration (INSDC). Resources provided by the DDBJ include traditional nucleotide sequence data released in the form of 27 316 452 entries or 16 876 791 557 base pairs (as of June 2012), and raw reads of new generation sequencers in the sequence read archive (SRA). A Japanese researcher published his own genome sequence via DDBJ-SRA on 31 July 2012. To cope with the ongoing genomic data deluge, in March 2012, our computer previous system was totally replaced by a commodity cluster-based system that boasts 122.5 TFlops of CPU capacity and 5 PB of storage space. During this upgrade, it was considered crucial to replace and refactor substantial portions of the DDBJ software systems as well. As a result of the replacement process, which took more than 2 years to perform, we have achieved significant improvements in system performance.

INTRODUCTION

The DNA Data Bank of Japan (DDBJ) is one of three databanks that constitute the DDBJ/ENA/GenBank International Nucleotide Sequence Database Collaboration (INSDC) (1–4), which was established through close collaboration with the European Bioinformatics Institute (EBI) and the US National Center for Biotechnology Information (NCBI). The DDBJ is dedicated to promotion, support, dissemination and use of biological data as public resources, and provides biological databases and accompanying analytic services geared for those purposes.

As part of the INSDC, DDBJ databases support the traditional assembled sequence archive known as 'Trad', the sequence read archive (SRA) for raw reads of new generation sequencers (5), and the BioProject and BioSample databases (6,7). Furthermore, DDBJ databases also support registration and auxiliary software systems.

The DDBJ is a division in the National Institute of Genetics (NIG) and is funded as a supercomputer centre with a 5-year replacement cycle. In 2012, the year the system was due for replacement, we took the opportunity to meet the dramatic increase in the demand of computational resources in genomics.

In this article, we report on submissions to DDBJ databases in 2012, provide background information on the new supercomputer system and discuss the progress made in refactoring of our software tools.

DDBJ ARCHIVAL DATABASES IN 2012

DDBJ traditional assembled sequence archive

Between July 2011 and June 2012, 15 243 000 entries and 12 270 462 217 base pairs (BP) were added to the DDBJ periodical release. This periodical release, also known as the INSD core traditional nucleotide flat files, does not include whole-genome shotgun (WGS) and third party annotation (TPA) files (8). The DDBJ contributed 17.8% of the entries and 12.0% of the base pairs added to the core nucleotide data of INSD in total. Most of the nucleotide data records provided to the DDBJ (97.8%) were submitted by Japanese researchers; the rest came from Korea (1.29%), China (0.57%), Taiwan (0.04%) and other countries and regions (0.31%). Furthermore, the DDBJ has continuously distributed sequence data in

*To whom correspondence should be addressed. Tel: +81 4 7136 3981; Fax: +81 4 7136 3975; Email: tt@nig.ac.jp
Correspondence may also be addressed to Yasukazu Nakamura. Tel: +81 55 981 6859; Fax: + 81 55 981 6886; Email: yanakamu@nig.ac.jp
Correspondence may also be addressed to Osamu Ogasawara. Tel: +81 55 981 5836; Fax: +81 55 981 5837; Email: oogasawa@nig.ac.jp

Table 1. List of large-scale data released by the DDBJ from July 2011 to June 2012

Type	Organism	Accession number (number of entries)
Genome	*A. digitifera*	scaffold CON: DF093604–DF097774 (4171 entries) WGS: BACK01000001–BACK01053640 (53 640 entries)
	Sake yeast (*S. cerevisiae* Kyokai no. 7)	scaffold CON: DG000037–DG000052 (14 entries) WGS: BABQ01000001–BABQ01000705 (705 entries) Mitochondrion: AP012028 (1 entry)
	Liver fluke (*C. sinensis*)	Phase 1 scaffold CON: DF126616–DF142827 (16 212 entries) WGS: BADR01000001–BADR01060778 (60 778 entries) Phase 2 WGS: BADR02000001–BADR02006190 (6190 entries) scaffold CON: DF142828-DF145382 (2555 entries)
	Hitomebore rice (*O. sativa* Japonica Group cv. Hitomebore)	scaffold CON: DG000053–DG000064 (12 entries) WGS: BACJ01000001–BACJ01064745 (64 745 entries)
	Eucaly (*E. camaldulensis*)	scaffold CON: DF097775–DF126446 (28 672 entries) WGS: BADO01000001–BADO01274001 (274 001 entries)
Full-length cDNA	Silkworm (*B. mori*)	AK377185–AK388575 (11 160 entries; 231 entries dropped)
	Pig (*S. scrofa*)	AK389169–AK401026(11 858 entries)
TSA	*B. braunii*	FX056085–FX112549 (56 465 entries)
EST	*B. braunii*	FY358876–FY368220 (9,345 entries)
	Sea squirt (*H. roretzi*)	FY844421–FY896670 (52 250 entries)
	Silkworm (*B. mori*)	FS724152–FS939542, FY736910–FY762881 (241 363 entries)
	Eucaly (*E. camaldulensis*)	FY782538–FY841121 (58 584 entries)
	Bread wheat (*T. aestivum*)	HX000001–HX201765, HX247045–HX257200 (211 921 entries)
	Honey bee (*A. m. carnica*)	HX282115–HX373155 (91 041 entries)
	Human (*H. sapiens*)	HY000001–HY377477 (377 477 entries)
	Asian Swallowtail (*P. xuthus*)	FY174038–FY210626 (36 589 entries)
	Common Mormon (*P. polytes*)	FY302525–FY358875 (56 351 entries)
MGA	Human (*H. sapiens*)	AEAAA0000001–AEAAA0026367, AEAAB0000001–AEAAB0012114, AEAAC0000001–AEAAC0021096, AEAAD0000001–AEAAD0024262, AEAAE0000001–AEAAE0023437, AEAAF0000001–AEAAF0030485, AEAAG0000001–AEAAG0021798, AEAAH0000001–AEAAH0040734, AEAAI0000001–AEAAI0029614, AEAAJ0000001–AEAAJ0030206 (260 113 entries)

published patent applications from the Japan Patent Office (JPO, http://www.jpo.go.jp) and the Korean Intellectual Property Office (KIPO, http://www.kipo.go.kr/en). The JPO transferred their data to the DDBJ directly, whereas the KIPO transferred their data through an arrangement with the Korean Bioinformation Center (KOBIC). A detailed statistical breakdown of the number of records is shown on the DDBJ homepage (http://www.ddbj.nig.ac.jp/breakdown_stats/prop_ent.html#prop_ent-table). In addition to the core nucleotide data, the DDBJ has also released a total of 2 646 472 WGS entries, 35 531 425 MGA entries, 704 TPA entries, 6374 TPA-WGS entries and 1272 TPA-CON entries, as of 27 August 2012.

Noteworthy large-scale data released from DDBJ are listed in Table 1. As can be seen in that table, the DDBJ has released the following: transcriptome shotgun assemblies (TSA) and expressed sequence tags (EST) of a green alga (*Botryococcus braunii*) submitted by the National Institute for Environmental Studies; *Acropora digitifera* genome submitted by the Okinawa Institute of Science and Technology; sea squirt (*Halocynthia roretzi*) EST submitted by the University of Tokushima; sake yeast (*Saccharomyces cerevisiae* Kyokai No. 7) genome submitted by the National Research Institute of Brewing; EST and full length silkworm cDNAs (*Bombyx mori*) submitted by the National Institute of Agrobiological Sciences; two phases of liver fluke (*Clonorchis sinensis*) draft genomes submitted by Sun Yat-sen University (China); Human 5′SAGE tags submitted by the University of Tokyo; Hitomebore rice (*Oryza sativa* Japonica Group cv. Hitomebore) genome submitted by the Iwate Biotechnology Research Center; full length cDNAs of pig (*Sus scrofa*) submitted by the National Institute of Agrobiological Sciences; eucaly (*Eucalyptus camaldulensis*) genome submitted by the Kazusa DNA Research Institute; wheat (*Triticum aestivum*) EST submitted by Yokohama City University; honeybee (*Apis mellifera carnica*) EST submitted by RIKEN; human EST submitted by RIKEN; EST of Asian Swallowtail (*Papilio xuthus*) and Common Mormon (*Papilio polytes*) submitted by the National Institute of Advanced Industrial Science and Technology (AIST).

Sequence read archive

A Japanese genetic researcher Masaru Tomita (Director of the Keio University Institute for Advanced Biosciences) published a full-genome sequence through DDBJ SRA under the accession number DRA000583. This was the first time a Japanese person had released their personal genome under his/her actual name.

BioProject and BioSample databases

With the advent of applications of sequencing devices, some sequenced samples may be analysed in different ways. Similarly, one project may submit multiple data sets to INSD. For the convenience of data submitters, the EBI and NCBI started to organize samples in the BioSample database and project information in the BioProject database (6,7). In late 2012, the DDBJ started to accept submissions to the BioProject database under the tripartite data exchange framework. To date, the DDBJ accepted >500 projects. Additionally, the DDBJ began preparations to join the BioSample framework developed by the NCBI and EBI. The DDBJ team is developing a submission system that encourages submission of samples that comply with the standards developed by the Genomic Standard Consortium (9). All those submissions will be integrated under the submission entrance portal.

SYSTEM RENEWAL

Outline of the new computer system

In March 2012, we completely replaced our computer system, which will be used for the three following purposes:

(i) to construct INSD and archiving nucleotide sequence data,
(ii) to provide search and analysis services for the sequence data, including large-scale computational annotation of genome sequences,
(iii) to provide computational resources to the researchers by providing a login portal to the computer system.

The new DDBJ computer system will be required to accommodate an ~1000-fold data increase for the same amount of funding. To achieve this, it was necessary to replace the current hardware and software with much more cost effective and high-performance models. Therefore, our new hardware system was designed from the concept stage as a commodity cluster based (Beowulf type) supercomputer system. Furthermore, even though their purchase was deemed sensible as part of efforts to implement a high reliability, large-scale, distributed system 5 years ago, it was also necessary to refactor substantial portions of our software because the previous systems were highly dependent on specific purposed proprietary hardware and middleware. These included the distributed database management systems, the full text keyword search system and the HTTP servers.

The theoretical peak performance of the CPUs in the new system is 122.5 TFlops (maximum performance Rmax is 82.90), which ranks our system as 280th in the Supercomputer Top 500 list (as of June 2012, http://www.top500.org/). The system also boasts 2 PB high-speed HDDs (Lustre file storage system, http://www.lustre.org/) for calculation, and 3 PB of massive array of idle disks (MAID) storage for data archiving. In addition, for computer systems aimed at large data analysis, it is important to ensure sufficient I/O interconnection bandwidth and a swift data transmission rate between calculation nodes and storage systems. Therefore, in our system, calculation nodes in the cluster system are interconnected with 40 Gbps InfiniBand via fat topology, which provides a total theoretical I/O bandwidth of 980 Gbps. This system also includes computers with huge memory (10 fat and TB and 2 TB medium nodes), which are especially useful for *de novo* assembly of deep sequenced data.

Taken together, the overall CPU performance of the new system is now approximately six times faster than the previous system, and its disk space is five times larger. Further system enhancements are planned in 2014, which will roughly double all aspects of the newly installed system. More specifically, peak performance will reach approximately 400 TFlops and the total size of the storage system will reach 12.5 PB, about 20 and 10 times the size of the previous system, respectively. The general increase in CPU performance is now 1.22 times/year (10), so the performance improvement for the new system significantly exceeds what would be expected based in this area alone.

The major improved aspects of the software system are as follows:

(i) The processing time for the entire DDBJ release data (about one hundred and forty million records) decreased from several days to 4–10 hr. Those improvements have enabled us to simplify the logic of daily data processing as a whole.
(ii) The insertion rate of flat file data to DDBJ GetEntry database system increased from 100 records per second to between 3000 and 10 000 records per second. This has helped us to more smoothly manage database history information dating back to the inauguration of the DDBJ.
(iii) Due to I/O bandwidth improvements to the storage system and the increased CPU performance, the processing rate of the basic local alignment search tool (BLAST) (11) is now from 10 to 100 times faster than was possible with the previous system.

Refactoring DDBJ services

Part of the replacement and upgrade process involves migration of our system to open source software. This is still in progress due to the immense size of the various elements. In this section, we will provide an overview of the current state of DDBJ services. Although some of the DDBJ service functions may seem less powerful than the previous versions, it is important to remember that we are now working to recover those functions and believe that full refactoring of the DDBJ system will result in overall improvements to our services in the near future.

(i) Keyword search system (ARSA)

The DDBJ keyword search system has been migrated from the proprietary Shunsaku search engine based on the SIGMA algorithm (12,13), which is an extension of the Aho–Corasick string matching algorithm (14), to the Apache Lucene Ver. 3.1. and Apache Solr Ver. 3.1 (http://lucene.apache.org/solr/tutorial.html) open source

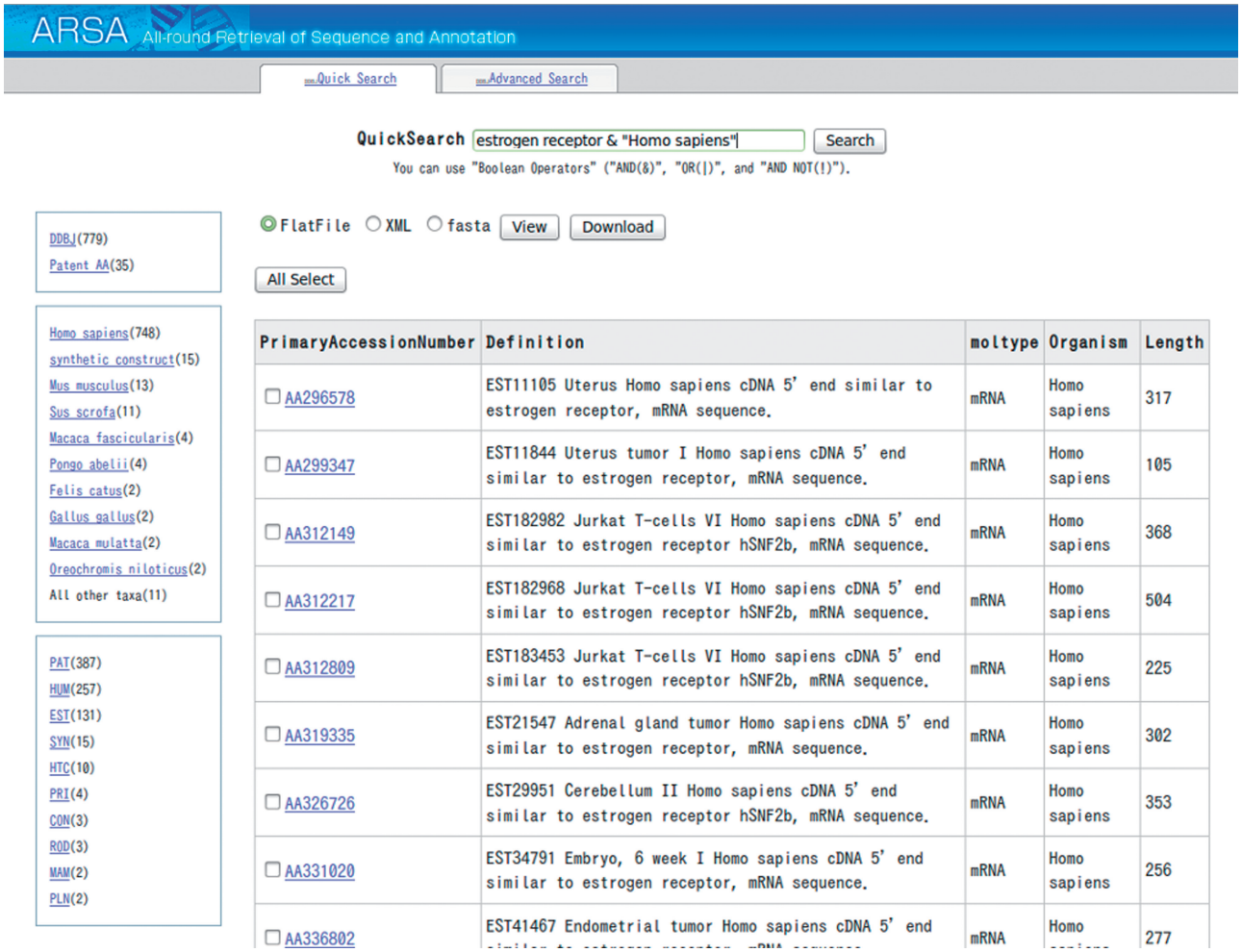

Figure 1. In the new DDBJ keyword search system, the web application was completely reimplementated to resolve the practical scalability limitations of the previous search engine. In the new system, not only normal divisions of INSD, but also EST, GSS, WGS have been made searchable.

software systems. These are the most widely used full text search engines based on the inverted index algorithm. The previous search engine operated in parallel for divided data stored on the memory, which made the search process fast and allowed input of complex queries with XPath syntax. This feature enabled our user to perform data retrieval by using terms in Feature/Qualifier identifiers of INSDC format. However, because of the practical scalability limitations, that system could not hold all DDBJ data, so EST, GSS, WGS, MGA data could not be searched.

That scalability problem has been successfully resolved and our new system is now (Nov, 2012) running on a small commodity cluster system. It is anticipated that we will soon be moving the system onto the new supercomputer (Figure 1).

(ii) WebBlast, TXSearch and ClustalW.

In the new system, the DDBJ restored a web-based BLAST tool with its original graphical user interface. The original package was NCBI BLAST 2.2.25 (11) and the prepared blast databases span all DDBJ's databases, including patents and the latest daily sequences. A selection box for INSDC division is also available.

TXSearch is a DDBJ-developed taxonomy database retrieval system that was unified by DDBJ, GenBank and ENA. This system is expected to be helpful as a reference for taxonomic names when researchers submit nucleotide sequences to the DDBJ. TXSearch utilizes the NCBI taxonomy database (15), and thus taxonomic accessions.

ClustalW, based on the original ClustalW 2.1 package (16), is also available on one of the DDBJ services. The DDBJ's ClustalW system contains original matrices developed at the National Institute of Genetics for genetics research.

(iii) DDBJ Read Annotation Pipeline

The DDBJ Read Annotation Pipeline (DDBJ Pipeline) is an NGS analytical system based on the NIG supercomputer (17). The DDBJ Pipeline began operation in 2009 and, since that time, >200 unique users worldwide have performed >4000 jobs using this system. In July 2012, the pipeline system was also migrated to the new NIG

supercomputer. The functional enhancement of DDBJ Pipeline covers four points, which will be explained below.

First, memory spaces for large genome *de novo* assembly were expanded up to 10 TB, which will be convenient for velvet users. Second, an RNA-seq *de novo* assembler was added. Third, mate-paired reads are supported for file upload. And finally, multiple query sets are now supported because the job number limitation for single users was abolished. This allows, for example, 20 DRA RUN files to be concurrently managed for data analysis.

FUTURE DIRECTION

As for software services, a beta version of the new sequence registration system has been tested, and we plan to develop an integration portal for new DDBJ registrations and other tools. Further development of cross search features for all DDBJ databases and support for large-scale NGS analytical systems are also crucial tasks. As for DDBJ services, the data scale in several search tools, and the Pipeline system itself, will be enlarged in near future.

ADDITIONAL INFORMATION

More information is available on the DDBJ website at http://www.ddbj.nig.ac.jp. News is delivered by really simple syndication (RSS), Twitter and mail magazines.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of all members of the DDBJ for their assistance in data collection, annotation and release, and software development. We are also thankful for the assistance of all advisors during the supercomputer replacement project.

FUNDING

Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) via a management expense grant for Inter-University Research Institute Corporation to the DDBJ; Grant-in-Aid for Scientific Research on Innovative Areas (Genome Science) to the DDBJ, SRA and DDBJ Pipeline (partial). Funding for open access charge: MEXT management expense grant to the DDBJ.

Conflict of interest statement. None declared.

REFERENCES

1. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, D38–D42.
2. Amid,C., Birney,E., Bower,L., Cerdeño-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Gibson,R., Goodgame,N., Hunter,C. *et al.* (2012) Major submissions tool developments at the European Nucleotide Archive. *Nucleic Acids Res.*, **40**, D43–D47.
3. Benson,D.A., Karsch-Mizrachi,I., Clark,K., Lipman,D.J., Ostell,J. and Sayers,E.W. (2012) GenBank. *Nucleic Acids Res.*, **40**, D48–D53.
4. Karsch-Mizrachi,I., Nakamura,Y., Cochrane,G. and International Nucleotide Sequence Database Collaboration (2012) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **40**, D33–D37.
5. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
6. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
7. Gostev,M., Faulconbridge,A., Brandizi,M., Fernandez-Banet,J., Sarkans,U., Brazma,A. and Parkinson,H. (2011) The BioSample Database (BioSD) at the European Bioinformatics Institute. *Nucleic Acids Res.*, **40**, D64–D70.
8. Cochrane,G., Bates,K., Apweiler,R., Tateno,Y., Mashima,J., Kosuge,T., Mizrachi,I.K., Schafer,S. and Fetchko,M. (2006) Evidence standards in experimental and inferential INSDC Third Party Annotation data. *OMICS*, **10**, 105–113.
9. Yilmaz,P., Kottmann,R., Field,D., Knight,R., Cole,J.R., Amaral-Zettler,L., Gilbert,J.A., Karsch-Mizrachi,I., Johnston,A., Cochrane,G. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.
10. Hennessy,J.L. and Patterson,D.A. (2012) *Computer Architecture: A Quantitative Approach*, 5th edn. Morgan Kaufmann Publishers, Burlington, MA.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Arikawa,S. (1981) One-way sequential search systems and their powers. *Bull. Math. Stat.*, **19**, 69–85.
13. Arikawa,S., Shinohara,T. and Takeya,S. (1989) SIGMA: a text database management system. *Berliners Informatik Tag*, 72–81.
14. Aho,A.V. and Corasick,M.J. (1975) Efficient string matching: an aid to bibliographic search. *Comm. ACM*, **18**, 333–340.
15. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. Kaminuma,E., Kosuge,T., Kodama,Y., Aono,H., Mashima,J., Gojobori,T., Sugawara,H., Ogasawara,O., Takagi,T., Okubo,K. *et al.* (2010) DDBJ progress report. *Nucleic Acids Res.*, **39**, D22–D27.